SST: Single-Stream Temporal Action Proposals

Implementation and Performance Analysis on the ActivityNet Benchmark

Ines Chami

Stanford University, Institute for Computational and Mathematical Engineering

Introduction

Video understanding is of particular interest nowadays because of the massive growth of video data available online. In order to understand and analyze these videos, scientist must develop algorithms that can localize and analyze important actions in videos.

Problem: We consider the problem of temporal action proposal in long videos that consists to retrieve temporal segments that are likely to contain an action of interest.

- Previous work on temporal proposal was mainly focused on sliding windows methods which can be very inefficient at test time.
- Traditional methods for action proposal operate on short sequences.

Objective: In this project we propose to implement the Single-Stream Temporal Action Proposals (SST) [1] model and compare it's performance to the DAPs approach on the ActivityNet dataset [2].

The SST Pipeline

The SST model was recently proposed for temporal action proposals. Given an input video, the model produces temporal intervals that are likely to contain actions of interest. The novelty of this model lies in the fact that the proposals are generated continuously, in a single forward pass and that it can operate on long input videos.

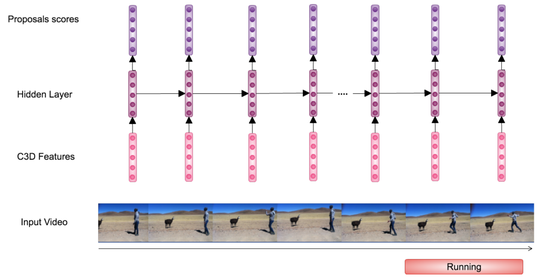

Figure 1: Schematic illustration of the SST architecture.

- **Visual Encoder**: Videos are first encoded using the top layer of a 3D convolution (C3D) network
- **Dense Generation of Training Examples**: Extraction of temporal segments of length W with stride s. This allows to generate more training examples and consider each timestep in different context during training.
- **GRU**: C3D features and then passed through a recurrent network that outputs proposal K scores for each timestep t: $\{\hat{s}_t{}^j\}_{j=1}^K$ where $\hat{s}_t{}^j$ is a score representing the likelihood that a proposal starts at time $t-j$ and ends at time t

Dataset and Evaluation Metrics

ActivityNet Dataset: We use the ActivityNet 200 [2]. It contains 19,994 videos (10,024 for training, for 4,926 for validation and 5,044 for testing). The C3D features dimensionality is reduced using PCA, resulting into 500-dimensional vectors.

Evaluation Metrics:Following the work of [1], we compute Recall@K for different values of K (between 100 and 1000) and for varying ious.

Figure 2: Data statistics for videos and proposals in the ActivityNet dataset, video duration distribution (left), distribution of proposal duration to video duration ratio (center), distribution of proposal duration (right)

Experiments and Results

The parameters K and W characterize the maximum proposal length and the maximum video length that can be captured by the SST model and must be judiciously selected.

(W,K)	Recall	FPS
(128,64)	20.37 %	441
(256,128)	32.43%	327
(512,256)	43.02%	206

Table 1: Recall@1000 for IoU=0.8 for the proposal generation task for varying values for K and W number of frames processed per second

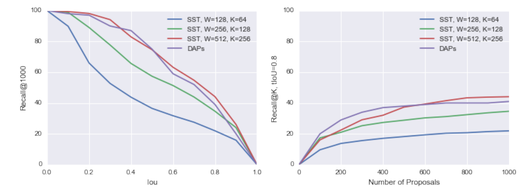

Figure 3: Recall@1000 as a function of IoU (left) and Recall@K for IoU=0.8 as a function of K, the number of proposals (right)

- Performance increases with W and K but so does computational complexity (FPS decreases): we must find a threshold
- The performance of the best SST model are better or comparable to those of the DAPs model
- SST operates in one pass and is therefore preferable to DAPs in terms of computational efficiency
- SST for $W = 512$ and $K = 228$ achieves better recall than DAPs for high IoU regime

Conclusions

- We implemented the SST model and tested it's performance on the Activitynet dataset
- We showed that SST could achieve state-of-the-art performance for the atemporal proposal generation task with significantly less computations needed.
- It is also possible to extend SST for more complex tasks such as video captioning.

References

[1] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.